DOI:10.2478/bile-2013-0015

Biometrical Letters
Vol. 50 (2013), No. 1, 27 - 38

Clustering of Symbolic Data based on Affinity Coefficient: Application to a Real Data Set

Áurea Sousa[1], Helena Bacelar-Nicolau[2], Fernando C. Nicolau[3], Osvaldo Silva[4]

[1]University of Azores, Department of Mathematics, CEEAplA, and CMATI,
9501-855-Ponta Delgada, Portugal, aurea@uac.pt
[2]University of Lisbon, Faculty of Psychology, Laboratory of Statistics and Data Analysis
1649-013-Lisboa, Portugal, and DataScience, hbacelar@fp.ul.pt
[3]New University of Lisbon, FCT, Department of Mathematics, 2829-516-Caparica, Portugal,
and DataScience, geral@datascience.org
[4]University of Azores, Department of Mathematics, CMATI, 9501-855-Ponta Delgada,
Portugal osilva@uac.pt

SUMMARY

In this paper, we illustrate an application of Ascendant Hierarchical Cluster Analysis (AHCA) to complex data taken from the literature (interval data), based on the standardized weighted generalized affinity coefficient, by the method of Wald and Wolfowitz. The probabilistic aggregation criteria used belong to a parametric family of methods under the probabilistic approach of AHCA, named *VL* methodology. Finally, we compare the results achieved using our approach with those obtained by other authors.

Key words: Ascendant Hierarchical Cluster Analysis, Symbolic Data, Interval Data, Affinity Coefficient, *VL* Methodology.

1. Introduction

The increasing use of databases, often large ones, in diverse areas of study makes it pertinent to summarize data in terms of their most relevant concepts. These concepts may be described by types of complex data, also known as symbolic data. In a symbolic data table, lines correspond to symbolic objects and columns to *symbolic variables*, which may assume not just one value, as usual, but multiple values, such as subsets of categories, intervals of the real axis, or frequency distributions. Furthermore, symbolic data tables may describe heterogeneous data and their cells may contain data of different types that can be weighted and linked by logical rules and taxonomies (Bock & Diday, 2000).

Relational databases are an important source of symbolic objects when we wish to study the properties of a set of data units. Symbolic data arise in a number of different ways (for example, as the result of aggregation of large data sets to obtain a data set of manageable size, or as a result of some scientific question(s) of interest).

Let $E= \{1,..., N\}$ be a set of data units described by p *interval* variables, $Y_1,...,Y_p$. A symbolic variable Y_j is regarded as an interval variable if for all $k \in E$ the subset $Y_j(k)$ is an interval of the real data set $\Re$. In this paper, we are dealing with this type of variable, often present in real data sets.

The aim of cluster analysis (of classical data as well as of symbolic data) is to build, from a (classical or generalized) data matrix ($N \times p$), a classification which is appropriate for a set E of data units (objects) or a set Y of variables, with the purpose of obtaining "homogenous" clusters of elements in a population Ω *or* E, so as to allow elements of the same cluster to present great similarity, whereas elements of different clusters will be much more different. Hierarchical methods yield a nested sequence of partitions of the elements to be classified. On the other hand, the partitional methods seek to obtain a single partition of the input data into a fixed number of clusters. The latter usually produce clusters by (locally) optimizing an adequacy criterion.

Many measures of proximity between symbolic objects have been proposed. An exhaustive review of some well-known measures of dissimilarity between symbolic objects is reported in Esposito et al. (2000). In this paper, we address only the case of clustering of symbolic data units described by variables whose values are intervals. Some dissimilarity coefficients for the particular case of interval data can be found in the literature (see e.g. Chavent and Lechevallier, 2002; Chavent et al., 2003; Souza and De Carvalho, 2004; De Carvalho et al., 2006*a*, 2006*b*).

Bacelar-Nicolau et al. (2009) use a similarity coefficient, namely, a generalized affinity coefficient (Matusita, 1951; Bacelar-Nicolau, 1980, 1988, 2000), for clustering data units described by interval variables, within the scope of hierarchical clustering of complex and heterogeneous data. In this paper, we again use the generalized affinity coefficient, as the basis of hierarchical clustering methods for interval data sets. Given the affinity similarity matrix, an interval data set can be classified through classical agglomerative hierarchical algorithms or probabilistic ones. The probabilistic aggregation criteria used (*AVL*, *AV1* and *AVB*) under the probabilistic approach of AHCA, named *VL* methodology (*V* for Validity, *L* for Linkage), resort essentially to probabilistic notions for the definition of the comparative functions. In fact, the *VL*-family is a set of agglomerative hierarchical clustering methods, based on the cumulative distribution function of basic similarity coefficients (Bacelar-Nicolau, 1980, 1988; Nicolau, 1983; Nicolau and Bacelar-Nicolau, 1998).

Section 2 contains the formula of the weighted generalized affinity coefficient for the case of interval variables. Section 3 contains the formula of the asymptotic standardized weighted generalized affinity coefficient, considering a permutational reference hypothesis R based on the limit theorem of Wald and Wolfowitz. In Section 4, we present the main results obtained with the application of Ascendant Hierarchical Cluster Analysis (AHCA) to interval data, based on the standardized weighted generalized affinity coefficient and on probabilistic aggregation criteria in the *VL* methodology. Finally, Section 5 contains some concluding considerations about the work and its results.

2. Weighted Generalized Affinity Coefficient for the Case of Interval Data

Based on the affinity coefficient between two discrete probability distributions as proposed by Matusita (1951), Bacelar-Nicolau (1980, 1988) introduced the affinity coefficient in Cluster Analysis as a basic similarity coefficient between the pairs of columns or lines of a data matrix, according to the set of elements that we wish to classify. Later on she extended that coefficient to different types of data, including complex data (symbolic data) and variables of mixed types (heterogeneous data), possibly with different weights (Bacelar-Nicolau, 2000;

Bacelar-Nicolau et al., 2009, 2010). The extension of the affinity coefficient to the case of symbolic data (in Symbolic Data Analysis) is called the weighted generalized affinity coefficient.

The weighted generalized affinity coefficient *a*(*k,k'*) between a pair of statistical data units *k, k'* $\in$ *D (k, k'=1,...,N),* may be defined as follows:

$$a(k,k') = \sum_{\mathrm{j}=1}^{\mathrm{p}} \pi_{\mathrm{j}} \cdot aff\left(k,k';j\right) = \sum\nolimits_{j=1}^{p} \pi_j \cdot \sum\nolimits_{\ell=1}^{m_j} \sqrt{\frac{x_{kj\ell}}{x_{kj\bullet}} \cdot \frac{x_{k'j\ell}}{x_{k'j\bullet}}} \qquad (1)$$

where: *aff (k,k';j)* is the generalized local affinity between *k* and *k'* over the *j-th* variable, m_j represents the number of modalities of a generalized sub-table associated with the *j-th* variable; x_{kjl} is a real non-negative value (a suitable adaptation of the formula (1) may be considered if real or frequency negative values appear) whose meaning depends on the type of the *j-th* variable (e.g. a discrete variable described by a frequency distribution or histogram, a binary vector or an interval variable);

$$x_{kj\bullet} = \sum\nolimits_{\ell=1}^{m_j} x_{kj\ell}\ ,\ x_{k'j\bullet} = \sum\nolimits_{\ell=1}^{m_j} x_{k'j\ell}$$

and π_j are weights such that $0 \leq \pi_j \leq 1,\ \Sigma\, \pi_j = 1.$ Either the local affinities or the whole weighted generalized affinity coefficient take values in the interval [0,1] and satisfy a set of proprieties which characterize affinity measurement as a robust similarity coefficient (e.g. Bacelar-Nicolau (2002), Bacelar-Nicolau et al. (2009)).

It should be emphasized that the weighted generalized affinity coefficient *a(k,k')* supports in a consistent way the use of Cluster Analysis models for statistical data units, when mixed and complex variable types are present in a database. In other words, the same coefficient – hence a unique algorithm – works for those variable types (Bacelar-Nicolau et al., 2009, 2010). However, we concentrate here on interval data. In the particular case of symbolic variables of interval type, Bacelar-Nicolau defined the weighted generalized affinity coefficient in the following way (for details, see Bacelar-Nicolau et al., 2009, 2010):

Let $E=\{1,\ldots,N\}$ be a set of data units described by p *interval* variables $Y_1,\ldots,Y_p$ and let Y_j be an interval variable, where j belongs to $\{1,\ldots,p\}$. Each cell $(k\ ,j)$ of the data matrix contains an interval $I_{kj}=[a_{kj}, b_{kj}]$ of the real axis,, with $k=1,\ldots,N$ and $j=1,\ldots,p$, and the k-th row $([a_{k1}, b_{k1}], \ldots, [a_{kp}, b_{kp}])$ describes the data unit k (see Table 1).

The weighted generalized affinity coefficient between a pair of data units k, k' ($k, k'=1,\ldots,N$) is given by:

$$a(k,k')=\sum_{j=1}^{p}\pi_j\cdot\frac{\left|I_{kj}\cap I_{k'j}\right|}{\sqrt{\left|I_{kj}\right|\cdot\left|I_{k'j}\right|}}, \qquad (2)$$

where $\left|I_{kj}\right|$ and $\left|I_{k'j}\right|$ symbolize, respectively, the ranges of the intervals I_{kj} and $I_{k'j}$.

It should be noted that formula (2) arises as a particular case of formula (1) when we are dealing with variables of interval type, as is demonstrated in Bacelar-Nicolau et al. (2009, 2010), from the decomposition of each interval into a suitable number of elementary intervals.

3. Asymptotic Standardized Weighted Generalized Affinity Coefficient

The values of a proximity matrix and clustering results are strongly affected by the modification of the scales of the variables. Usually, some standardization must be performed prior to the clustering process in order to attain an 'objective' or 'scale-invariant' result. Under this condition, standardization greatly improves the performance of the clustering method (De Carvalho et al., 2006*a*).

On the assumption of a permutational reference hypothesis R based on the limit theorem of Wald and Wolfowitz (Fraser, 1975), the random variable associated with $aff(k,k';j)$ has asymptotic normal distribution, and a standardized weighted generalized affinity coefficient $a_{WW}(k,k')$ by the method of Wald and

Wolfowitz may be used, instead of *a(k,k')* (see e.g. Bacelar-Nicolau, 1988; Bacelar-Nicolau et al., 2009; Bacelar-Nicolau et al., 2010).

Table 1. Symbolic data table (interval data)

	Y_1	$\cdots$	Y_j	$\cdots$	Y_{1p}
1	$[a_{11},b_{11}]$	$\cdots$	$[a_{1j},b_{1j}]$	$\cdots$	$[a_{1p},b_{1p}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$[a_{k1},b_{k1}]$	$\cdots$	$[a_{kj},b_{kj}]$	$\cdots$	$[a_{kp},b_{kp}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	$[a_{N1},b_{N1}]$	$\cdots$	$[a_{Nj},b_{Nj}]$	$\cdots$	$[a_{Np},b_{Np}]$

The standardized weighted generalized affinity coefficient by the method of Wald and Wolfowitz (Bacelar-Nicolau et al., 2010) is given by the formula:

$$a_{WW}(k,k') = a^{*}(k,k') = \sum_{j=1}^{p} \pi_j .aff^{*}_{WW}(k,k';j)\,, \tag{3}$$

where the local asymptotic normal affinity coefficient $aff^{*}_{WW}(k,k';j)$ also satisfies the main properties of a similarity coefficient. Furthermore, the coefficient a_{WW}(*k,k'*) allows us to define a probabilistic coefficient within the *VL* methodology, under the approach begun by Lerman (1972, 1981) and developed by Bacelar-Nicolau (e.g. 1980, 1987, 1988) and Nicolau (e.g. 1983, 1998).

4. Case Study: Freshwater fish data set (Ecotoxicology data set)

The Freshwater fish data set (Ecotoxicology data set) *set* consists of a set of 12 species of freshwater fish described by 13 interval variables. Table 2 shows part of the corresponding data matrix (De Carvalho et al., 2006*b*); the complete data matrix is available in the SODAS (*Symbolic Official Data Analysis System)* Software. According to De Carvalho et al. (2006*b*), several studies carried out in French Guyana indicated abnormal levels of mercury contamination in some Amerindian populations. This contamination was connected to their consumption of contaminated freshwater fish. In order to study this phenomenon, the data set mentioned above was collected by researchers from the LEESA (*Laboratoire d'Ecophysiologie et d' Ecotoxicologie des Systèmes Aquatiques*) laboratory.

Table 2. Freshwater fish data set

Interval	Individuals	/Labels		
Variable	Ageneiosusbrevifili	Cynodongibbus	…	Myleusrubripinis
Length	[22.5:35.5]	[19:32]	…	[12.3:18]
Weight	[170:625]	[77:359]	…	[80:275]
Muscle	[1425:5043]	[2393:8737]	…	[8:35]
Intestine	[333:2980.06]	[0:2653]	…	[0:0]
Stomach	[0:1761.1]	[478.34:10860.7]	…	[10.76:41.93]
Gills	[393.71:853.1]	[354.22:1976.38]	…	[0:9.45]
Liver	[642:7105.77]	[2684.83:43014]	…	[190.12:394.52]
Kidneys	[0:3969.05]	[1437.82:27514.6]	…	[72.3:112.54]
Liver/Muscle	[0.45:1.41]	[1.12:4.92]	…	[7.12:30.35]
Kidneys/Muscle	[0:2.02]	[0.6:3.24]	…	[2.42:10.23]
Gills/Muscle	[0.15:0.3]	[0.15:0.24]	…	[0:0.85]
Intestine/Muscle	[0.23:0.63]	[0:0.5]	…	[0:0]
Stomach/Muscle	[0:0.55]	[0.2:1.24]	…	[0.31:4.33]

The symbolic objects (species) were grouped into four *a priori* clusters according to diet. The *a priori* classification is as follows (De Carvalho et al., 2006*b*):

Carnivorous:
1-Ageneiosusbrevifili/C 2- Cynodongibbus/C 3- Hopliasaimara/C
4-Potamotrygonhystrix/C

Detrivorous:
7-Dorasmicropoeus/D 8- Platydorascostatus/D 9- Psedoancistrusbarbatus/D
10-Semaprochilodusvari/D

Omnivorous:
5-Leporinusfasciatus/O 6- Leporinusfrederici/O

Herbivorous:
11-Acnodonoligacanthus/H 12- Myleusrubripinis/H

In order to apply the affinity coefficient $a_{WW}(k,k')$**,** with equal weights $(\pi_j=1/p)$, a transformed data matrix was computed, according to that described in Bacelar-Nicolau et al. (2009, 2010). Each interval variable (generalized column) gave a sub-table with a suitable number of columns corresponding to a set of elementary intervals. The affinity coefficient was combined with three probabilistic aggregation criteria, *AVL*, *AV1*, and *AVB* (Bacelar-Nicolau, 1988; Nicolau, 1983; Nicolau and Bacelar-Nicolau, 1998).

Table 3 represents the similarity matrix, and Figures 1 and 2 show the dendrograms associated with the *AVL* and *AV1/AVB* aggregation criteria respectively.

Table 3. Similarity matrix: coefficient $a_{WW}(k,k')$

	1/C	*2/C*	*3/C*	*4/C*	*5/O*	*6/O*	*7/D*	*8/D*	*9/D*	*10/D*	*11/H*	*12/H*
1/C	1.00											
2/C	1.26	1.00										
3/C	2.02	1.53	1.00									
4/C	0.51	0.15	1.17	1.00								
5/O	0.26	-0.16	-0.36	-0.32	1.00							
6/O	0.30	-0.34	-0.36	0.25	0.38	1.00						
7/D	0.92	0.96	0.90	1.49	0.31	0.06	1.00					
8/D	0.46	0.36	0.24	1.68	0.39	0.28	2.01	1.00				
9/D	-0.76	-0.30	-0.93	0.39	-0.14	-0.21	-0.15	0.30	1.00			
10/D	0.59	0.33	0.08	0.93	0.12	0.22	0.98	1.15	-0.19	1.00		
11/H	-1.22	-0.71	-0.98	-0.12	-0.33	0.03	-0.05	0.54	1.41	-0.11	1.00	
12/H	-0.88	-0.41	-0.71	-0.49	0.40	-0.55	-0.38	0.16	0.30	-0.13	0.98	1.00

In this section, we obtain a comparison between the results of AHCA based on the generalized affinity coefficient, and those obtained by De Carvalho et al. (2006*b*) based on the dynamic clustering algorithm (partitional method) considering different distances between vectors of intervals: adaptive Hausdorff distance, non-adaptive Hausdorff distance, one-component adaptive city-block distance and non-adaptive city-block distance. These results are also compared with the *a priori* classification.

Table 4 and Figures 1 and 2 show that the four main clusters obtained from the application of the asymptotic standardized generalized affinity coefficient $a_{WW}(k,k')$ combined with the *AVL*, *AV1* and *AVB* methods are all the same, and they may explain the partition given by the L1 and Hausdorff adaptive methods considered in De Carvalho et al. (2006*b)*: all the methods find the *a priori* classification, except for the two elements *4/C* and *9/D*. The L1 and Hausdorff non-adaptive methods give a different result. Furthermore, the cluster {*9/D 11/H 12/H*} was found in all partitions into four clusters obtained from all of the clustering methods.

Table 4. Clustering results for the Freshwater fish data set

Method	*Cluster 1*	*Cluster 2*	*Cluster 3*	*Cluster 4*
AHCA (a_{WW}(k,k′) + AVL) *AHCA (a_{WW}(k,k′) + AV1)* *AHCA (a_{WW}(k,k′)+ AVB)*	*5/O 6/O*	*9/D 11/H 12/H*	*1/C 2/C 3/C*	*4/C 7/D 8/D 10/D*
L1 (non –adaptive) *Hausdorff (non –adaptive)*	*1/C 4/C 7/D 8/D 10/D*	*2/C*	*3/C*	*5/O 6/O 9/D 11/H 12/H*
L1 (adaptive) *Hausdorff (adaptive)*	*5/O 6/O*	*9/D 11/H 12/H*	*1/C 2/C 3/C*	*4/C 7/D 8/D 10/D*

```
                  Levels 1 a 11
1/C   --*--*
           |--------*
3/C   --*--*        |-----------*
                    |           |
2/C   --*-----------*           |
                                |-----*
4/C   --*-----*                 |     |
              |--------*        |     |
7/D   --*     |        |        |     |
        |-----*        |--------*     |
8/D   --*              |              |
                       |              |
10/D  --*--------------*              |
                                      |
5/O   --*--------------------*        |
                             |-----*  |
6/O   --*--------------------*     |  |
                                   |--*
9/D   --*--------*                 |
                 |--------*        |
11/H  --*--------*        |--------*
                          |
12/H  --*-----------------*
```

Figure 1. Dendrogram obtained with AVL

```
                  levels 1 a 11
1/C   --*--*
           |--------*
3/C   --*--*        |--------*
                    |        |
2/C   --*-----------*        |
                             |--------*
4/C   --*-----*              |        |
              |--------*     |        |
7/D   --*     |        |     |        |
        |-----*        |-----*        |
8/D   --*              |              |
                       |              |
10/D  --*--------------*              |
                                      |
5/O   --*-----------------------*     |
                                |--*  |
6/O   --*-----------------------*  |  |
                                   |--*
9/D   --*--------*                 |
                 |--------*        |
11/H  --*--------*        |--------*
                          |
12/H  --*-----------------*
```

Figure 2. Dendrogram obtained with AV1/AVB

5. Concluding Remarks

In this paper, we have presented some applied results obtained from Ascendant Hierarchical Cluster Analysis (AHCA) of symbolic objects described by interval data, in order to illustrate the effectiveness of AHCA based on the standardized weighted generalized affinity coefficient by the method of Wald and Wolfowitz, for symbolic data.

We have analyzed a real data set and compared the results obtained using this coefficient with the results obtained by other authors with different methods. The Ascendant Hierarchical Cluster Analysis (AHCA) methods based on the affinity coefficient are shown to be quite robust even in the case of a small sample. In fact, similar dendrograms were obtained considering the three probabilistic aggregation criteria (*AVL*, *AV1*, and *AVB),* and all of the three hierarchical methods found the *a priori* classification, as did the best non-hierarchical methods from De Carvalho et al.

REFERENCES

Bacelar-Nicolau H. (1980): Contributions to the Study of Comparison Coefficients in Cluster Analysis, PhD Th. (in Portuguese), Univ. Lisbon.

Bacelar-Nicolau H. (1987): On the Distribution Equivalence in Cluster Analysis, Proc. of the NATO ASI on Pattern Recognition Theory and Applications, Springer-Verlag, New York, 1987: 73-79.

Bacelar-Nicolau H. (1988): Two Probabilistic Models for Classification of Variables in Frequency Tables. In: Classification and Related Methods of Data Analysis, H.-H. Bock (ed.), North Holland: Elsevier Sciences Publishers B.V.: 181-186.

Bacelar-Nicolau H. (2000): The Affinity Coefficient. In: Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, H.-H. Bock and E. Diday (Eds.), Berlin: Springer-Verlag: 160-165.

Bacelar-Nicolau H. (2002): On the Generalised Affinity Coefficient for Complex Data. Biocybernetics and Biomedical Engineering 22(1): 31-42.

Bacelar-Nicolau H., Nicolau F.C., Sousa A., Bacelar-Nicolau L. (2009): Measuring Similarity of Complex and Heterogeneous Data in Clustering of Large Data Sets, Biocybernetics and Biomedical Engineering 29(2): 9-18.

Bacelar-Nicolau H., Nicolau F.C., Sousa A., Bacelar-Nicolau L. (2010): Clustering Complex Heterogeneous Data Using a Probabilistic Approach. Proceedings of Stochastic Modeling Techniques and Data Analysis International Conference (SMTDA2010), Chania Crete Greece, 8-11 June 2010 – published on the CD Proceedings of SMTDA2010 (electronic publication).

Bock H.-H., Diday E. (2000): Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Series: Studies in Classification, Data Analysis, and Knowledge Organization, Berlin: Springer-Verlag.

Chavent M., Lechevallier Y. (2002): Dynamical Clustering Algorithm of Interval Data: Optimization of an Adequacy Criterion Based on Hausdorff Distance. In: Classification, Clustering, and Data Analysis, K. Jajuga, A. Sokolowski, H.-H. Bock (Eds.), Berlin: Springer-Verlag: 53-60.

Chavent M., De Carvalho F.A.T., Lechevallier Y., Verde R. (2003): Trois Nouvelles Méthodes de Classification Automatique de Données Symboliques de type intervalle, Revue de Statistique Appliquée, tome 51(4): 5-29.

De Carvalho F.A.T., Brito P., Bock H-H. (2006*a*): Dynamic Clustering for Interval Data Based on L_2 Distance. Computational Statistics 21(2).

De Carvalho F.A.T., Souza R.M.C.R. de, Chavent M., Lechevallier Y. (2006*b*): Adaptive Hausorff Distances and Dynamic Clustering of Symbolic Interval Data. Pattern Recognition Letters 27(3).

Esposito F., Malerba D., Tamma V. (2000): Dissimilarity Measures for Symbolic Objects, In: Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, H.-H. Bock and E. Diday (Eds.), Berlin: Springer-Verlag: 165-185.

Fraser D.A.S. (1975): Non Parametric Methods in Statistics. Chapman and Hall.

Lerman I.C. (1972): Étude Distributionelle de Statistiques de Proximité entre Structures Algébriques Finies du Même Type: Apllication à la Classification Automatique. Cahiers du B.U.R.O., 19, Paris.

Lerman I.C. (1981): Classification et Analyse Ordinale des Données, Paris: Dunod.

Matusita K. (1951): On the theory of Statistical Decision Functions, Ann. Instit. Stat. Math. III: 1-30.

Nicolau F.C. (1983): Cluster Analysis and Distribution Function. Methods of Operations Research 45: 431-433.

Nicolau F.C.m, Bacelar-Nicolau H. (1998): Some Trends in the Classification of Variables. In: Data Science, Classification, and Related Methods, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, Y. Baba (Eds.), Springer-Verlag: 89-98.

Nicolau F.C. (1983): Cluster Analysis and Distribution Function. Methods of Operations Research 45: 431-433.

Souza R.M.C.R. de, De Carvalho F.A.T. (2004): Clustering of interval data Based on City-Block distances, Pattern Recognition Letters 25: 353-365.